DOCUMENT RESUME

ED 173 348 TM 007 426

AUTHOR Lennon, Roger T.
TITLE Perspective on Intelligence Testing.
PUB DATE 30 Mar 78
NOTE 9p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)
JOURNAL CIT NCME Measurement in Education; v9 n2 Spr 1978

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Culture Free Tests; Elementary Secondary Education; Intelligence Factors; *Intelligence Tests; Labeling (of Persons); *Opinions; Teacher Attitudes; *Test Bias; *Testing Problems; *Test Interpretation; Test Reliability

ABSTRACT

We should seek an updated perspective on intelligence testing because it is useful to reevaluate any practice that has long become institutionalized, especially one that is subject to severe criticism. Cultural bias is the most prominent criticism. Testing proponents contend that intelligence tests reflect skills and knowledge emphasized in school (although these may be culturally influenced), and that test makers are using sophisticated techniques to detect biased items. Some object to intelligence tests because they measure a narrow range of abilities; few test makers can identify other abilities which would increase predictive ability as efficiently. Test makers contend that teachers can and should develop and review expectations of individual pupil progress, based on intelligence test results; furthermore, the vast majority of teachers and adults feel that the tests do not measure inborn intelligence alone. Some critics discount results because they fluctuate over time; but a study with 2,000 students retested on the Otis-Lennon Mental Ability Test, after three and six years, challenged this view. Human concerns center around labeling and quantification of individuals' ability. In conclusion, the original need to group students still prevails; intelligence tests have improved; and their discontinuation would harm both disadvantaged and mainstream groups. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173348

Perspective on Intelligence Testing*

Roger T. Lennon
The Psychological Corporation

ABOUT THIS REPORT

Roger T. Lennon

In this issue, Dr. Roger Lennon presents an enlightening view of the "forest" of intelligence testing unobstructed by the "trees" of intelligence tests. The issues are presented cogently and objectively, understanding that few are more qualified to discuss this current and still controversial topic.

Dr. Lennon needs no introduction to the measurement community. His pursuit and attainment of excellence has been documented over a career spanning approximately 40 years of professional life. He has served in myriad capacities to both NCME, APA, and AERA. He has authored and co-authored numerous contributions to the professional literature.

Dr. Lennon has recently resigned his position as officer of Harcourt Brace Jovanovich and chairman of the Psychological Corporation to assume his new position as Associate to the Chairman of Harcourt Brace Jovanovich in measurement.

PSR

TM 007426

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
I. J. Lehmann
NCME

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Perspective, the dictionary tells us, has as one its meanings the "capacity to view things in their true relations or relative importance," and it is in this sense that I interpret the term in the title of this address. Achievement of a perspective on intelligence testing that might lay claim to being final or definitive is surely an overly ambitious goal, for this or perhaps any occasion; but I hope that even a modest essay in that direction may not be without benefit.

Perspective is achieved in various ways — as by standing back, by viewing elements from several vantage points, by considering these elements with a certain detached attitude rather than as through the eyes of one imtimately involved as participant or actor in the matters under view. For obvious reasons it has not been easy for me to achieve this degree of remoteness or disinterestedness, but I have made the effort — with what success you must judge: it has not been total.

This audience needs no definition of the kinds of tests that I shall be talking about under the rubric "intelligence tests" — even though many of them now mask their identity under supposedly more attractive aliases. I shall have in mind those tests that profess to measure a more or less general ability, or a very small number of such abilities, assumed to underlie or condition performance across a considerable range of tasks primarily intellectual in character. The tests may yield either a single estimate of this hypothetical ability, or may seek to assess it in two or three modalities such as verbal and non-verbal; or with attention to two major sub-domains, most commonly labeled *verbal* and *quantitative.* Typical of the tests I have in mind, though not exhaustive of the category, are the *Wechsler Intelligence Scales* and the *Stanford-Binet,* among individual measures, and, among group instruments, the *Thorndike-Hagen Cognitive Abilities Test, California Test of Mental Maturity: Short-Form, Otis-Lennon Mental Ability Test, Short Test of Educational Ability, Tests of General Ability,* and even *Primary Mental Abilities Test,* insofar as this, despite its factor-analytic character, nevertheless also yields a general measure of intelligence. I shall not detain us here with a consideration of whether the abilities measured by any of these

*Invited address, National Council on Measurement in Education, Toronto, Ontario, March 30, 1978.

tests are properly called "intelligence," nor about the differences in the constructs measured by the several instruments, insofar as these may be inferred from the respective test content. I will only observe that study of the descriptions of what each test purports to do and to measure, and analysis of the content elements of the several tests, warrants discussing them as a group, even though the observations may apply with greater or less precision to each of the tests named and others like them. I shall also limit the discussion to the use of these tests in connection with elementary and secondary education, simply as a matter of keeping it within reasonable compass, and close to where I think your interests lie.

A moment ago, I alluded in a fashion some of you may have thought facetious to the practice of calling new tests in this category, or new editions of them, something other than "intelligence" tests. I do not wish to leave the impression that all the changed designations are merely cosmetic or huckstering devices; they do reflect more modest representations for the various instruments, or efforts to make more precise the nature of what is being assessed. (Parenthetically I report that the *Otis-Lennon* which through its title now professes to be a measure of *mental* ability, will, in its next edition, identify itself as Otis-Lennon *School* Ability Test.) Whatever the labels, I will refer unashamedly to the tests here concerning us as "intelligence tests" — not even prefacing that terms with "so-called"; you will not be confused.

Why should we seek a perspective on intelligence testing? My answer is this: It is useful to reconsider any practice that has long become institutionalized, that may well be taken for granted, and that may be persisting out of unthinking habit, or inertia. One ought periodically to consider whether the conditions and the needs which originally gave rise to the adoption of any practice or procedure still obtain, whether advances in knowledge would suggest superior or more efficient ways of meeting those needs or addressing those conditions if they still persist, whether unintended consequences of a practice, either not foreseen or not present in earlier times, now bring into question the cost-benefit equations associated with the practice; and so on.

It seems particularly timely to ask ourselves questions of this kind with respect to intelligence testing in connection with education. No one in this audience needs to be told that all testing has for the last few years been the subject of intense scrutiny, frequently hostile; and of all the types of testing, intelligence testing has been the object of the most severe, not to say virulent, criticism and attack. The dimensions or themes of this attack are now clearly marked out, and any effort to establish a perspective on intelligence testing must surely attend to these elements of the scene and strive to assay their "relative importance."

The criticisms, for the most part, are not new. Also, again for the most part, they are serious, though, in all candor, my biases impel me to see some of them as trivial, frivolous, ill-informed or worse, as I have elsewhere declared. They deserve to be, as they have been, addressed seriously by measurement persons and educators generally. Some have found one or several of the criticisms to be sufficently persuasive to prompt the banning of the use of group intelligence tests, as did New York City, for example, the state of California, and a few other jurisdictions. Others, finding the criticisms lacking in merit or insufficiently persuasive, have formally elected to continue the use of the tests. Many are simply confused about the matter, proceeding perhaps more tentatively than they might have a decade ago. There are, to the best of my knowledge, no dependable data on changes in the extent of use of intelligence tests that would signal major shifts in the attitudes or practices of users. I can report, with respect to the use of intelligence tests published by The Psychological Corporation, no evidence of widespread diminution in their use. I estimate, again on the basis of less than adequate information, that about two-thirds of school districts in this country now make use of intelligence tests on some regular or p.ogrammatic basis. You will be interested to hear that in a survey of testing attitudes among NCME

NCME

measurement in education

Vol. 9, No. 2 Spring, 1978

Paul S. Redelheim, Editor

Advisory Committee:
Sidney E. Benton, North Georgia College
Carmen Finley, American Institutes for Research
George S. Ingebo, Portland Public Schools
Anthony J. Polemeni, New York City Board of Education
Warren W. Willingham, Educational Testing Service

Measurement in Education is a series of special reports published four times a year in October, January, March, and May by the National Council on Measurement in Education. These reports are concerned with the practical implications of measurement and related research and their applications to educational problems of individuals, institutions, and systems. The emphasis is upon uses of measurement rather than technical or theoretical issues. Subscription rate: $5.00 per year; single copies 75¢ each in quantities of 25 or more, or $1.50 for a single issue.

Second-class postage paid at East Lansing, MI 48823 and at Cleveland, OH

Address Editorial correspondence to:

Dr. Paul S. Redelheim
George Peabody College for Teachers
Box 158
Nashville, TN 37203

Address Business correspondence to:

Dr. Irvin J. Lehmann
202 S. Kedzie Hall
Michigan State University
East Lansing, MI 48824

members (and reported on by Stetz at this meeting), 9% of those responding predicted increased attention to intelligence testing ten years from now, 42% look for no change from the present, and 42% predict less attention. So you can sense the uncertainty even among us about the prospects for intelligence testing.

Within the confines of the time available, I shall comment on what seem to me the most serious of the questions and criticisms that are currently being raised about intelligence tests. Particularly, I shall address the following concerns: cultural bias, nature of abilities measured, limited utility and redundancy of results, labeling, misunderstanding and misuse of tests results, technical and psychometric concerns, dependability of results, human considerations in testing. This list, while a full platter for a luncheon address, is not exhaustive, but most of the other criticisms that are advanced derive from, or are variants in some fashion of the ones here enumerated. Obviously I cannot go into a thorough discussion of the merits of each of the criticisms. Articles, if not books, can be and have been written with respect to each of the points, and by now the *pro* and *con* positions are more than adequately spelled out in literature. I will attempt to do no more than suggest briefly the conflicting considerations with respect to each of the issues, as a kind of backdrop to help achieve the perspective that is the goal.

Cultural bias

By far the most prominent criticism of intelligence tests, and underlying and informing many of the other criticisms, is that such tests are culturally biased. The heart of the cultural bias charge is that most of the commonly used intellignece tests, either individual or group, are unfair to certain subgroups in the population, most notably minority groups, because their content is inappropriate for the minority examinee. In this view, typical test content is perceived as excessively reflective of white middle- or upper-class values, typically developed by authors who have little or no understanding of the differences in experience between middle-class groups and those from other segments of society. Such content is violative of the assumption underlying some interpretations of test scores, namely, that all examinees will have had reasonably similar experience and background before coming to the test. Thus, the systematic differences regularly found in group performance cannot be accepted as evidence of true differences in the constructs presumably being measured by the tests, but rather must be considered as consequent upon differing experiential backgrounds. Exception is also taken to the exclusion or under-representation of minority groups in the norming samples for the tests, leading to the assertion that it is not possible to evaluate the performance of an individual by relating it to the performance of a norms group appreciably different in important charateristics. As a result of the unfairness of the test, the abilities of the disadvantaged examinees are underestimated, with resulting harm to the educational and other opportunities available to them.

Proponents of these tests, on the other hand, point out that the critical consideration in evaluating the fairness of the tests should be the ability of the scores to predict performance on relevant criteria for all groups — fairness can be evaluated only in relation to use; that there need be no concern for the sources of variance on the test scores, as long as no inferences are drawn about some presumed potential, much less inherited capacity; that the test is accurately descriptive of the here-and-now performance of the subject on tasks that can be shown to be significant with respect to near-term or even longer-term learnings.

The values that the school strives to cultivate are, of course, culturally saturated. Those skills and knowledges whose presence in the tests is deemed objectionable are precisely those emphasized in the school. Thus, if one is interested in assessing a pupil's likelihood of succeeding in the learning of such material, it is sensible that the predictive measures similarly emphasize such material even though it may be culturally influenced. In fact, if the experiential backgrounds of certain groups do not equip their members with those competences that facilitate learning of school-set tasks, then it surely must be counted a virtue rather than a defect of a predictor that it be sensitive to the consequences of such background differences. It may even be argued that no device so efficiently captures those consequences of cultural differences that are important for schooling as does the intelligence test. If background factors have, in fact, contributed to the creation of handicaps to learning, surely it is better to know this than not to know it in the planning of instructional programs.

The literature abounds with technical definitions and ways of assessing bias in tests, as Dr. Gardner informed us in his presidential address yesterday.* The concept of bias is by no means a simple one, so that it is not possible to declare categorically that a test is biased or not biased. Furthermore, existence of bias in one or another of the technical senses does not necessarily mean that a test will result in unfair treatment of various subgroups of examinees.

Regardless of one's opinions on the existence or extent of bias, by whatever definition, in intelligence tests, there are measures that can be taken and, in fact, have been taken to improve the situation very much short of giving up the tests. Test makers generally are sensitive to matters of appropriateness of test content for various groups in the population, and most have moved to insure the representation of varied cultural groups and minorities in the development and review of test content. Most test makers, moreover, are now utilizing more sophisticated techniques for analysis of item performance to increase the likelihood of detection and elimination of items that may perform improperly differentially for various subgroups; I cite, for example, studies such as one being reported on today by Scheuneman. And, finally, most norms groups for tests developed in recent years have included proper representation of minority groups proportionate to their presence in the general population.

Nature of abilities measured

Some object to present intelligence tests on the grounds that they are too narrowly conceived, that they provide only a partial, and perhaps distorted, measure of the totality of an examinee's intellectual functioning. The tests are perceived as giving inordinate emphasis to verbal skills, to abstract thinking, and to tasks not perceived as meaningful by certain supgroups of the population. To be sure, the content of most intelligence tests, from Binet to the present, has been drawn heavily

*Dr. Eric Gardner's presidential address on bias will appear as Volume 9, #3 of Measurement in Education.

from about a dozen types of tasks: vocabulary, general information, analogical reasoning, series or sequence manipulation, perceptual acuity, spatial abilities, quantitative skills, classification, syllogistical reasoning — a familiar litany to you. It is not difficult to name types of intellectual activity that are missing from any of the commonly used tests, even broadly based ones such as Wechsler or Stanford-Binet. One has only to consider the Guilford structure-of-intellect-model, for example, and note how many of its cells are not reflected in existing tests, to know that the sample of abilities represented is indeed limited. The qu estion is whether it is excessively restricted for the purposes for which the instruments are used, and this is largely an empirical matter, a matter of cost-benefit relating testing time to adequacy of prediction or diagnosis. For the prediction of scholastic success in elementary and secondary grades, as measured by either standardized achievement tests or teacher grades, the commonly used intelligence or scholastic aptitude tests do a reasonably efficient job. Few test makers have been able to identify other types of tasks or test content which would, in reasonable time, add appreciably to the predictive effectiveness of the present instruments. There should not be, and ordinarily there is not, any pretense that the results of general ability tests constitute an adequate indexing, much less mapping, of the whole of a subject's intellectual ability. There are probably few now on either side of the intelligence testing argument who would object in pinciple to the extension of the range of abilities tapped, provided it could be demonstrated that such extension was warranted in terms of testing time and improved prediction, either against the existing scholastic criteria or other measures of school adjustment, or against new types of educational experiences that might be more appropriate avenues for socially beneficial cultivation of such new sets of abilities.

Limited utility and redundancy of results

Related to the concern for the limited scope of the abilities measured by existing tests is the criticism that the information yielded by these tests is too undifferentiated and general to serve as a useful guide for planning instruction; or, that the information is redundant — duplicative of what is otherwise known about pupils. This criticism runs to the heart of the use of such tests for pupil classification and grouping and is, in my opinion, possibly the most serious of the questions or objections raised to the use of intelligence tests.

Most makers of these tests would, I believe, say that such criticism is based upon an excessive expectation of the way in which the results might be used for instructional intervention. It is certainly true that the results of intelligence tests do not provide the classroom teacher with a prescription as to suitable content, sequence, methodology, or mode of presentation most likely to be effective day by day in the learning progress of the examinee. The tests are not designed to furnish this kind of information. It does not follow, however, that they are therefore of no significance or utility to the teacher attempting to accommodate instruction to pupil needs. One need only study the masses of data accumulated in many places over long periods of time on the relations between intelligence test scores and achievement scores or teacher grades, whether these data are collected simutaneously or whether the achievement scores are obtained a year or more years after the intelligence test scores, to be persuaded that whatever the intelligence tests are measuring, it bears a substantial relationship, current or prospective, to the actual learnings of pupils. Ironically, these relationships have sometimes been so strong as to lead to the criticism that the two types of tests really were not measuring anything different — the old Truman Kelley "jingle-jangle fallacy." The correlations are high; but who should be surprised that bright children on the whole do well in school, and dull ones less so? There is still ample independent variance to justify exploration of significant differences between the two. The notion that measured achievement by itself tells us all that it is useful to know about a learner's scholastic equipment is at least arguable.

We can assert with considerable confidence how persons at various levels on intelligence tests will perform in the academic disciplines at various grade levels: we know, for example, what range of arithmetic or spelling or reading achievement will be attained by, e.g., fifth-grade pupils of various IQ levels and we are in the process of developing even more specific information on the relation of intelligence test scores to task performance. Shall we say that such information is of no value in establishing appropriate instructional goals for these varying ability levels, in making some tentative judgements about the likely learning pace of these pubils, about the probable degree of abstraction with which they can cope successfully? In the opinion of most intelligence test makers, these are proper and helpful inferences to draw, with whatever degree of tentativeness, from the test results, and while they would represent only the most rudimentary beginnings in the planning of appropriate instructional intervention, they are surely important first steps.

It is, of course, precisely at this point that critics take violent exception. They see such activity as evidence of the harmful consequences, namely, the provision of diminished or limited educational programs and opportunities to those persons appearing on the basis of test scores to be lacking in the abilities required to deal with the programs. The diminished expectations, say the critics, all too often lead to diminished attainments. Indeed they may, but they need not. Teachers must, in fact, entertain expectations for individual learners if they are to make any pretense at all to providing appropriate instruction for individual pupils. The issue is not whether teachers should have expectations. They must. The issue is rather how such expectation may best be developed and whether test information contributes to more accurate or less accurate expectations; and finally, the adequacy of processes established to provide for continuing review of expectations.

Some makers of aptitude and intelligence tests have taken the position that the teacher ought to regard it as her task to upset the predictions, to provide the kinds of instructional programs that will permit leaners to exceed the level of attainment characteristic of pupils of similar intelligence. We have paid too little attention to the fact that at whatever level of intelligence the range of achievement associated with it is large: as mentioned above, there *is* independent variance and pupils even in what might be considered fairly low reaches of intelligence in some instances manifest very creditable achievement. What do we know about the characteristics or conditions associated with this kind of successful performance? Not as much as we should. We tend to think of such pupils that they apply themselves with diligence

or have personality characteristics that result in their being looked on with favor by teachers while they are grading them, which may of course be true. But there may be other characteristics of the instructional situation that contribute.

We know very little about aptitude-treatment interaction in the case of general intelligence test results and the usual achievement test results. We have a shortage of predictor instruments known to be of specific utility in pointing the way toward appropriate instructional strategies and interventions, and the prospects are not great that we shall add largely to our store of this kind of knowledge in the near future: the accumulation of this type of data is a formidable, almost discouragingly difficult task. But, if anything, this lack of knowledge on how to adapt the instructional program to varied aptitude characteristics of learners ought to make us even more hesitant about abandoning such tools as we have, granted their utility in this respect is limited.

Labeling

The kinds of classification of learners implied in the previous section have led thoughtful observers to express concern that examinees who perform poorly on intelligence tests incur the risk of being categorized and identified as "slow learners," "academic risks," "mentally retarded," or by some similarly pejorative designation, with consequent damage to their self-esteem, to the regard in which they are held by others, and to educational and employment opportunities. This categorization and labeling comes about as a result of a need to group and classify individuals for administrative purposes, given the impracticality in the ordinary school situation of treating each pupil on a totally individual basis. The classifications and labels are tools of convenience. That they can be, and perhaps often have been, hurtful, both to self-image and to reputation, cannot be gainsaid. Efforts to substitute kindlier euphemisms have been tried, with indifferent success. As long as individual differences exist and some individuals have conspicuously less of a given ability or attribute than others, and these differences imply or call for different educational treatments, it is hard to see how classification and consequent labelings can be avoided — and this whether the classifications are made on the basis of test results or otherwise. The most useful advice would appear to be that the classifications and labelings be made with full awareness of the possibilities for error in arriving at the judgments, with all the precautions against the hurtful consequences of such labeling that concerned and ingenious educators can devise, and with not only a readiness but a dedication to continuous monitoring of the bases for classification, and willingness to change whenever the evidence warrants.

Misunderstanding and misuse of test results

We are told that, whatever the mertis or defects of the test themselves, their results are likely to be misused because teachers and parents have so little understanding, or so much postitive misunderstanding, of what it is that tests measure and the kinds of inferences the test results support. The point is made that teachers or parents tend to attach excessive significance to the results of tests, repose too great confidence in them, and impute to them an immutability that is assumed to be associated with belief in the constitutional or hereditary character of whatever is measured by the tests. Some critics have it that hundreds of thousands of pupils are victims of such erroneous conceptions.

We may well agree that teachers receive inadequate training at either a pre-service or in-service level with respect to measurement generally with respect to the nature and use of intelligence test particularly. NCME's own inquiries have documented this. However, it appears unlikely — to me at least — that teachers characteristically entertain the kinds of attitudes or misconceptions about the meaning of intelligence test results implied in the criticsm above. The median age of teachers is in the early 30's; most of them, therefore, have had their pre-professional and professional training within the past 15 years. It would be difficult to find a textbook in psychology, educational psychology, or psychological testing in use during that time that would have led any reader to suppose that intelligence tests were measures of innate intelligence; in fact, the textbooks of this era, if anything, have emphasized an environmentalist point of view. Insofar as the textbooks have influenced teacher attitudes, they would surely have induced a cautious and tentative approach to the interpretation and application of results of these tests. Nor can one find in the manuals for any of the currently widely-used intelligence tests any assertions that would give rise to or reinforce an excessively hereditarian view of what is measured.

As long ago as 1967, the Goslin study, *Teachers and Testing,* told us about the views of teachers on what intelligence test measured. To the question, "Do you think intelligence tests measure the intelligence a person is born with or what he has learned?" put to a sample of teachers, only 1.1% responded that they thought the tests measured only inborn intelligence. The overwhelming majority endorsed the view that the tests measured inborn intelligence and learned knowledge in varying proportions. A sample of American adults when asked the same question overwhelmingly chose responses signifying a belief that intelligence tests measured learned knowledge as well as inborn abilities. It does not follow that teachers or adults generally have altogether correct notions about the significance of test results, but at least they are probably not guilty, in general, of the special kinds of misconceptions ascribed to them. The declaration in a *Time* article in December 1977 is right: "The day is long past when the IQ was revered as some sort of magic number, affixed during childhood as a indelible, immutable badge of mental prowess or dullness."

My judgment is that we do an injustice to the vast majority of teachers to suppose that they entertain naive and erroneous ideas about the nature of what is measured by typical intelligence tests. On the whole it seems to me that they have a very sensible understanding of the influences that contribute to test performance, that they do not worry overmuch about the resources of pupil performance at a given time, but take the test results as currently descriptive of an important constellation of knowledges, skills, and abilities. Popular articles provide us with their share of horror stories about gross misinterpretations and foolish or harmful actions taken by teachers and institutions on the basis of such misconceptions. There is no doubt there have been bad espisodes; but our ignorance about such matters is great. It would be a service to all to have dependable information on the prevalence or rarity of such misuses of intelligence test results, and the harmful consequences flowing from

REPORTS AVAILABLE

Back issues of *Measurement in Education* are available for 1.50 each per single issue, and for 75¢ each in quantities of 25 or more for a single issue.

Vol. 1,	No. 1	*Helping Teachers Use Tests* by Robert L. Thorndike
	No. 2	*Interpreting Achievement Profiles — Uses and Warnings* by Eric F. Gardner
	No. 3	*Mastery Learning and Mastery Testing* by Samuel T. Mayo
	No. 4	*On Reporting Test Results to Community Groups* by Alden W. Badal & Edwin P. Larsen
Vol. 2,	No. 1	*National Assessment Says* by Frank B. Womer
	No. 2	*The PLAN System for Individualizing Education* by John C. Flanagan
	No. 3	*Measurement Aspects of Performance Contracting* by Richard E. Schutz
	No. 4	*The History of Grading Practices* by Louise Witmer Cureton
Vol. 3,	No. 1	*Using Your Achievement Test Score Reports* by Edwin Gary Joselyn & Jack C. Merwin
	No. 2	*An Item Analysis Service for Teachers* by Willard G. Warrington
	No. 3	*On the Reliability of Ratings of Essay Examinations* by William E. Coffman
	No. 4	*Criterion-Referenced Testing in the Classroom* by Peter W. Airasian and George F. Madaus
Vol. 4,	No. 1	*Goals and Objectives in Planning and Evaluation: A Second Generation* by Victor W. Doherty and Walter E. Hathaway
	No. 2	*Career Maturity* by John O. Crites
	No. 3	*Assessing Educational Achievement in the Affective Domain* by Ralph W. Tyler
	No. 4	*The National Test-Equating Study in Reading (The Anchor Test Study)* by Richard M. Jaeger
Vol. 5,	No. 1	*The Tangled Web* by Fred F. Harcleroad
	No. 2	*A Moratorium? What Kind?* by William E. Coffman
	No. 3	*Evaluators, Educators, and the Publics: A Detente?* by William A. Mehrens
	No. 4	*Shall We Get Rid of Grades?* by Robert L. Ebel
Vol. 6,	No. 1	*On Evaluating a Project: Some Practical Suggestions* by John W. Wick
	No. 2	*Measuring Reading Achievement: A Case for Criterion — Referenced Testing and Accountability* by S. Jay Samuels and Glenace E. Edwall
	No. 3	*Security in a Citywide Testing Program* by Anthony J. Polemeni
	No. 4	*Dear Mama: Why don't they love me anymore?* by Thomas J. Fitzgibbon
Vol. 7,	No. 1	*The Discrepancy Evaluation Model (Part I)* by Andres Steinmetz
	No. 2	*The Discrepancy Evaluation Model (Part II)* by Andres Steinmetz
	No. 3	*What Graduate and Professional School Students Think About Admission Tests* by Letnard L. Baird
	No. 4	*The Paradox of Education Testing* by Robert L. Ebel
Vol. 8,	No. 1	*FACE UP: A Framework for Assessing Career Educators Utilization Programs* by Bruce W. Tuckman
	No. 2	*Grade Equivalent Scores* by Gary Echternacht and *If Not Grade Equivalent Scores — Then What?* by Jeanne M. Plas
	No. 3	*Criticisms of Testing: How Mean is the Median?* by Henry S. Dyer
	No. 4	*The Superintendent and Testing: Implications for the Curriculum* by Herbert C. Rudman
Vol. 9,	No. 1	*A Consumer's Guide to Criterion — Referenced Test Item Statistics* by Ronald C. Berk

them. Meanwhile, everyone must applaud whatever increased and improved training can be provided with respect to understanding and use of test of all kinds.

Technical and psychometric concerns

One theme in the current criticism of intelligence tests echoes concerns that have been present from the earliest years of such tests. These are concerns for the technical adequacy of the tests as measuring instruments. Disaffection is voiced with respect to such matters as the choice of types of test content; quality of individual test items; manipulation of item responses to generate scores; normative systems provided for the interpretation of the scores; similarities and differences between measurement in the physical sciences and in the behavioral sciences; matters of scaling, and other psychometric issues.

Those of you who saw the TV show, "The Myth of the IQ," a couple of years ago, remember the citation of selected items from various intelligence tests; the items clearly chosen to impress the lay observer as foolish, biased, erroneous, or otherwise defective. Articles in popular and even some semi-professional media similarly criticize individual items on familiar grounds: no right answer, more than one right answer, unfairness to unconventional types of thinkers, ambiguity, dependence on factors other than those supposedly being measured, etc. Test makers respond as you would expect: instruments, like any human products, are not perfect; given the enormous number of test items in all the published test, some poor items find their way into such tests despite maintenance of a generally high level of item quality through sophisticated editorial, review, and item analysis activities. Unless it is demonstrated that such items make up a signficant part of a test, their existence does not vitiate its usefulness. No one would deny that every effort should be made to keep their number to a minimum; and the general level of item excellence is certainly a proper basis for choosing among tests. Test makers regard cirticisms of individual items, whether from friends of testing or opponents, as legitimate types of criticism, each to be evaluated on the merits.

Criticisms directed at such matters as the homogeneity of scales, the additive properties of responses, equality of units, and the like, address issues well covered in measurement literature for 40 years or more — a circumstance, incidentally, with which not all making these criticisms seem familiar. Their treatment in recent attacks on intelligence tests does not impress me as having advanced either the level or the content of discourse about intelligence test methodology. Informed cirticisms on technical grounds are welcome, for they are addressible — that is, they may be considered in accordance with generally accepted criteria, with the expectation that competent judges would come to consensus about them on the basis of a reasonable evidence.

Among criticisms of this type we must include those directed at IQ itself — both as concept and as metric. The eminently sensible Henry Dyer has termed the IQ a "statistical monstrosity." This was perhaps a fair characterization of the original ratio IQ's; it seems excessively harsh with respect to IQ's as now derived, which are merely standard scores. A persuasive case can be made for elimination of this term, on the grounds that it now carries, in professional and lay minds alike, an insupportable freighting of emotional and otherwise

irrelevant connotations. But it is sensible to wonder whether it can that easily now be exorcised from the language, or whether the terms invented to replace it will be more accurately interpreted.

Dependability of intelligence test results

A special technical criticism runs to the dependability or reliability of test results. Some critics tend to discount these results on the grounds that they fluctuate so much from time to time for individuals as to render them of very limited usefulness in the planning of instruction, in grouping, or in other applications. This criticism is directed most frequently to the results of group ability measures, in the probably mistaken belief that results of individual intelligence tests are more reliable; published reliability data indicate that differences in this respect between individual measures and the better group measures are slight.

What can be asserted with respect to short- or long-term stability of intelligence test results? While there have been many studies of short-term fluctuations in individual IQ's over the years, there have been relatively few long-term studies. One such study was undertaken by Robertson and Mitchell of the results on *Otis-Lennon Mental Ability Test.* A sample of 2,321 pupils in three school systems who participated in the 1966 national standardization program for *Otis-Lennon* while they were in grades 1 through 6 were retested three years later when they were in grades 4 through 9, and again three years later when they were in grades 7 through 12. The correlation between first-test IQ's and second, three-year later, IQ's, was .83; the median value of the correlations for the separate grade groups was .86. The correlation between the second-test results and the third-test results, three years later than the second, was .89; the median value for the correlation for separate grade groups was .90. Finally, even more impressively, the correlation between first-test results and third-test results, with a six-year interval between testings, was .82, again with a median value of .83 for the separate grade groups. As would be expected, tests administered in grade 1 and grade 2 tended to correlate least with tests in later years, but even these correlations were high. A smaller sample of these same pupils was again retested three years later, thus nine years after initial testing; the resulting correlations, for pupils initially tested in grades 1, 2 and 3 respectively, were .69, .73, and .81. Given a median reported alternate-form reliability, for testings within a two-week period, of .90, these long-term stability coefficients have to be seen as attesting to a remarkable degree of consistency over time in whatever attribute is being measured by the test, at least under ordinary school conditions.

Not only is the rank order of performance stable as revealed in these correlations; the mean IQ changed only 2 points during the three years between first and second testing, less than one-half point between the second and third testing and, thus, only about 2-1/2 points over the six-year interval between first and third tests. Over the first three-year interval, 72% of pupils had IQ changes of less than 10 points, over the second three-year interval 83% had IQ changes of less than 10 points. These data persuade me, at least, that whatever set of abilities is being tapped by *Otis-Lennon,* it is not an ephemeral or transitory characteristic at least from the time an individual enters school. In fact, it is hard to think of many other human characteristics that would reveal similar stability over time. From this standpoint, the IQ deserves to be given serious consideration in educational planning.

Human considerations in testing

There continue to be some who feel that it is offensive or an affront to human dignity to strive to reduce the richness of a human person, or even of an attribute such as intelligence, to a number, or a set of numbers. Self-identified humanists see such quantification as tending toward depersonalization, loss of respect for the individual as an individual, a homogenization of regard and treatment for persons. Not so, say the test makers, some of whom also regard themselves as "humanists, " or at least not as insensitive wretches. To them, there seems nothing inhumane about seeking some measure of precision in the ways in which we describe human beings and their characteristics, nor in improving the acccuracy and communicabilty of our information about them, which is the function of a test score. Neither, this latter group would say, is there any way in which knowledge of human beings can be systematized and codified, manipulated and analyzed, except through its reduction to numerical terms. We could not even begin to study many of the problems that presently concern both friends and foes of tests, such as the intellectual consequences of economic deprivation, the effectiveness of various types of educational programs, the access to educational opportunities at all levels for various social groups, to name a few, without having information that is expressible in quantitative terms. It betokens little confidence in the good sense and good will of teachers, guidance counselors, psychologist and others, to imply that their valuing and treatment of a pupil or client is so fragile as to give way when an estimate of ability is provided as a number rather than in grosser descriptive terms.

I am aware that this paper so far has been largely an "on-the-one-hand — but on-the-other" presentation — and deliberately so, for I wanted to address you more as expositor of the current scene than as advocate. But I should like to avail myself of a closing minute or two to go on record once again with my personal biases in the matter.

You may take it for granted that I would not be engaged in a revision of the *Otis-Lennon* tests if I believed that this type of test had outlived its usefulness. I have considered the kinds of questions posed at the outset of this talk — do the needs and conditions which gave rise to the introduction of intelligence tests and their early rapid spread in the schools still obtain? Have advances in scientific knowledge about human abilities or the learning process rendered these tests obsolete? Have we become aware of unfortunate side effects that argue forcefully for the curtailment of use of such tests? and so on.

After reflection on such issues, I now espouse a credo that includes the following affirmations:

1. The original needs and conditions still prevail; the needs for classification and grouping of learners, however ambiguous or even negative the research on the effects of grouping; for better diagnosis and understanding of the nature of learning disabilities; for identification of talent; for setting realistic goals and establishing appropriate paces — expectations, to use the sensitive word; and these needs are better served with the availability of intelligence test

results than in the absence of such information.

2. Existing instruments have not been outpaced by advances in our knowledge of human intellectual functioning, or in our understanding of learning and instruction; the instruments have improved steadily, if not dramatically, over time, and their fundamental design and content is still relevant.
3. There are problems in the use of these tests with culturally deviant groups that we have been insufficiently sensitive to, but the proposed response, namely the discontinuation of the use of the tests, would work far more harm than good, not only for the mainstream group, with whom their use has been relatively freer from criticism, but with the very groups now feeling disadvantaged by the tests.

I have said nothing so far about the role — an indispensable one in my veiw — of such tests in the description of groups when such description is essential, as for program evaluation, accountability studies, research enterprises where matching of groups on mental ability or some such construct is called for. Nor have I dwelt on the role of instrumentation in investigations of intellectual functioning; in this field, as in so many others, advances in knowledge only become possible with improved instrumentation, and vice versa. If our whole armamentarium of intelligence tests were destroyed tomorrow, and the technology underlying their construction, I have no doubt it would have to be reinvented if our understanding of human behavior is to flourish.

The mental ability test may not be, as Lee Cronbach termed it some years ago, psychology's greatest single contribution to the welfare of society, but I am willing to devote my energies and resources to its cultivation.

References

Gardner, Eric F. *Bias.* Presidential Address, National Council on Measurement in Education, Toronto, Ontario, March 29, 1978.

Goslin, David A. *Teachers and Testing.* Russell Sage Foundation, New York, 1967.

Robertson, Gary T. and Blythe C. Mitchell. Stability of Mental Ability Test Performance over Three- and Six-Year Intervals. Unpublished paper. 1975. Available from The Psychological Corporation.

Scheuneman, Janice. *Ethnic Group Bias in Intelligence Test Items.* Paper presented at American Educational Research Association meeting, Toronto, Ontario, March 1978.

Stetz, Frank P. and Michael D. Beck. *A Survey of Opinions Concerning Users of Educational Tests.* Paper presented at National Council on Measurement in Education meeting, Toronto, Ontario, March 30, 1978.

Time. Whatever Happened to Genius? December 17, 1977.

NCME NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

202 S. Kedzie Hall
Michigan State University
East Lansing, Michigan 48824

Second class postage paid
at East Lansing, Mich.
48823 and
additional mailing points.